局所探索法による熱力学的DNA配列設計の改良 (理論計算機科学の深化 : 新たな計算世界観を求めて)

AUTHOR(S):

川下, 優; 小野, 廣隆; 定兼, 邦彦; 山下, 雅史

CITATION:

川下, 優 ...[et al]. 局所探索法による熱力学的DNA配列設計の改良 (理論計算機科学の深化 : 新たな計算世界観を求めて). 数理解析研究所講究録 2008, 1599: 27-34

ISSUE DATE:

2008-05

URL:

http://hdl.handle.net/2433/81799

数理解析研究所講究録
第1599巻 2008年 27-34

局所探索法による熱力学的DNA配列設計の改良

川下 優 * 小野 廣隆 † 定兼 邦彦 † 山下 雅史 †
* 九州大学大学院システム情報科学府 † 九州大学大学院システム情報科学研究院

1 はじめに

近年，DNA分子からなる塩基配列を利用したナノ技術・ナノコンピューティングが注目されている．DNA分子はワトソン・クリック相補性に基づいた結合・乖離反応を起こすが，これらの反応は自律的・並列的に起きるため，処理速度やエネルギー効率の面からその有用性が期待されている．また，分子の微小性から莫大な情報収納量が期待されている．塩基配列集合とワトソン・クリック相補性を利用することで，Adleman [1] は，生化学実験において有向ハミルトンパス問題を解くことに成功した．また，上述の情報格納を目指した分子メモリ [9] なども考案されている．

これらの技術の多くは，塩基配列を集合として利用している．この時，塩基配列集合はワトソン・クリック相補性を考慮した制約を満たす必要があると考えられている．また，集合中の配列数は各技術における資源数となるため，集合サイズは大きいことが望まれる．このため，上述の条件を満たす塩基配列集合の設計手法が必要とされている．

筆者らは過去に，局所探索法を基にした塩基配列集合の設計手法を提案し，既存の研究より良い結果を得ることに成功している [10, 11]．特に，[11] においては，塩基配列集合の制約として最小自由エネルギーと呼ばれる指標を用いた制約を採用した．しかし，提案手法による最小自由エネルギーを用いた配列設計には，問題点がある．最小自由エネルギー計算には多くの計算時間が必要であり，これが配列設計に必要な時間を増大させている．そこで，配列設計における最小自由エネルギー計算を軽減し，配列設計に必要な時間を削減することを目指す．

2 準備

2.1 塩基配列とワトソン・クリック相補性

塩基配列は生体高分子であり，A,T,G,Cのアルファベットを用いて表現する塩基から構成される．また，塩基配列は方向性を持つ一本鎖であり，その両端は5′, 3′のどちらかで表現される．塩基配列を構成する塩基数は配列長と呼ばれる．ここで，配列長 n の塩基配列は $s = s_1 s_2 \cdots s_n$ と表現され，$s \in \{\mathrm{A,T,G,C}\}^n$ となる (図 1)．

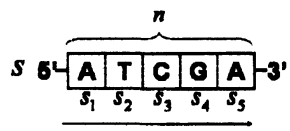

図 1: 塩基配列の図

また，塩基配列はワトソン・クリック相補性と呼ばれる特性を持つ．これは，4種類の塩基のうち，A-T間，G-C間でのみ水素結合が生じるというものである．このとき，塩基配列の構造上，二本の塩基配列は逆方向でなければならない．特に二本の塩基配列中のすべての塩基において塩基対が生じる場合，相補の関係にあるといい，二本の配列の一方を主配列とすると，もう一方を相補配列と呼ぶ．ここで，相補性に基づき $\overline{\mathrm{A}} = \mathrm{T}, \overline{\mathrm{T}} = \mathrm{A}, \overline{\mathrm{G}} = \mathrm{C}, \overline{\mathrm{C}} = \mathrm{G}$ と表すと，s の相補配列を $\overline{s} = \overline{s_n}\ \overline{s_{n-1}} \cdots \overline{s_1}$ と表現できる．このとき，A-T間，G-C間で常に水素結合が生じ

るとは限らない．また，一本鎖内で水素結合が生じる場合もある．

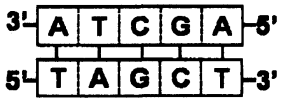

図 2: 相補関係での結合

2.2 塩基配列と形態

前述したように，塩基配列はワトソン・クリック相補性による水素結合を持つ．そこで，塩基配列は生じた水素結合に従った**配列形態**となる．このとき，A-T，G-C の組合せは多数存在するので，同じ塩基配列を与えても，配列形態は多数存在する．例えば，図 3 と図 4 はともに同じ塩基配列であるが，生じている水素結合は異なるため，異なる形態であると考える．

図 3: 二本鎖形態

図 4: 二本鎖形態

また，各形態は塩基配列と水素結合からなる**ループ構造**に分解できる．

2.3 自由エネルギー

塩基配列は，その形態により**自由エネルギー**と呼ばれる値が定められる．同一の塩基配列においても，形態が異なる場合には，自由エネルギーの値も異なる．自由エネルギーは最大値が 0 の実数で表現され，値が低い形態ほど安定することが知られている．このため，塩基配列は自由エネルギーが低い形態で安定する．

自由エネルギーの値は生化学実験より得られたものであり，各形態のエネルギーは，その形態を構成する各ループ構造の持つエネルギーの和により近似される．

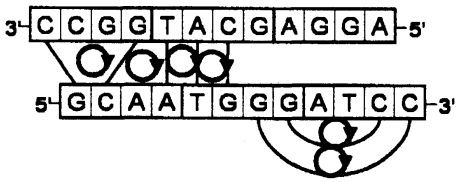

図 5: 自由エネルギー概念の図

特に，塩基配列が与えられたときに，取り得る自由エネルギーの中で最小のものを，**最小自由エネルギー** (Minimum Free Energy: **MFE**) と呼ぶ．最小自由エネルギーは，塩基配列により一意に定められる．また，最小自由エネルギーを取る形態が最も安定した形態となる．最小自由エネルギーは動的計画法を用いることで，二本の配列長 n_1, n_2 のときには，$O((n_1+n_2)^3)$ で求めることができる [2, 13, 20]．動的計画法で用いる表を示したものが図 6 である．表の縦軸と横軸は二本の塩基配列を繋げたものに対応する．また，図の●位置の値は，●の右上に存在する表の値を参照することで決定できる．

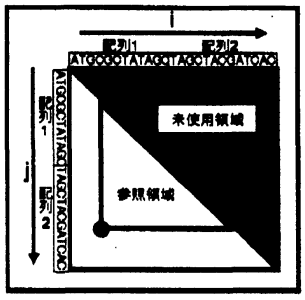

図 6: MFE 計算における DP 表

2.4 局所探索法

厳密な最適解を見つけることが極めて困難な (NP 困難などの) 組合せ最適化問題を解く際に，最適性の保証は無くとも精度が十分に高い似解が求まれば良いとされる場合がしばしば存在する．このような際に，簡易な方針にもかかわら

ず有効と知られているのが発見的手法の一つである**局所探索法**である．

組合せ最適問題は，**解** x，**目的関数** f とし，$f(x)$ を最小化（最大化）する問題と定式化できる．このとき，x に少しの変化を加えることで得られる解集合を**近傍**と呼び，$N(x)$ と表す．すべての $x' \in N(x)$ において $f(x) \leq f(x')$（もしくは $f(x) \geq f(x')$）を満たすとき，x を近傍 $N(x)$ における**局所最適解**という．局所探索法は，この局所最適解を求める解法である．ある解から近傍内の解を生成するために加える操作を**近傍操作**という．また，局所最適解は多数存在することが多いという特徴がある．

局所探索法の基本的な戦略は以下の通りである：

(1) 初期解 x を選択する．

(2) x の近傍 $N(x)$ の内を探索する．

(3) 改善解 x' が (2) で見付かった場合，$x := x'$ として (2) へ．そうでなければ，局所最適解として x を返す．

一般に，近傍中には改善解が複数存在し，近傍中をどのような順序で調べ，どのような改善解に移動するのかについては，様々な戦略がある．これを**移動戦略**といい，代表的なものとしては，近傍中で改善解が見つかり次第移動する**即時戦略**がある．

3 配列集合設計問題

塩基配列設計の研究において，配列集合設計問題が考えられており，既存の研究ではこの問題を取り扱っているものが多い．この配列集合設計問題とは，Adleman の実験 [1] に用いられた塩基配列集合に求められる条件を簡易化したものとなっている．本予稿においても，この問題を取り扱うこととする．

配列集合設計問題は，配列長 n，集合サイズ m である配列集合 S を設計する問題である．通常，n の値は比較的小さいものが対象となっている．このとき，S には以下の二点が要求される．

(a) $\forall s \in S$ は $\bar{s}$ 以外と結合して安定しない．

(b) m は大きい．

(a) の条件を満たすための制約はさまざま考案されている．多くの制約は，組合せ的制約と熱力学的制約に分類することができる．組合せ的制約は主にハミング距離を利用するものが多く [3, 4, 5, 6, 8, 10, 12, 16, 17]，熱力学的制約は最小自由エネルギーを利用するものが多い [6, 11, 14, 15, 17]．組合せ的制約は，熱力学的制約を近似的に表したものと捉えられている．このため，組合せ的制約は熱力学的制約より簡易ではあるが，(a) の条件を満たすための指標としての精度は熱力学的制約に劣る．

既存研究の多くにおいては簡易である組合せ的制約を取り扱っている．しかしながら，精度の高い熱力学的制約を用いることが近年重要視されてきており，熱力学的制約を利用した研究も行われるようになってきた．本予稿においては，熱力学的制約を用いた場合について考察する．

3.1 制約

二本の配列 s, s' が取る最小自由エネルギーを $\Delta G(s, s')$ で表し，τ を制約定数とするとき，上記 (a) の条件は，以下の制約として表現することができる．

(1) $\Delta G_{ww}(S) \stackrel{\text{def}}{=} \min_{s,s' \in S}\{\Delta G(s, s')\} \geq \tau$

(2) $\Delta G_{wc}(S) \stackrel{\text{def}}{=} \min_{s,s' \in S, s \neq s'}\{\Delta G(s, \overline{s'})\} \geq \tau$

(3) $\Delta G_{cc}(S) \stackrel{\text{def}}{=} \min_{s,s' \in S}\{\Delta G(\overline{s}, \overline{s'})\} \geq \tau$

以上の制約について考察する．

これらの制約について考える場合，配列設計問題は以下のように記述することができる：

入力 配列長 n，配列数 m，及び制約定数 τ．

出力 $\Delta G_{ww}(S) \geq \tau$，$\Delta G_{wc}(S) \geq \tau$，$\Delta G_{cc}(S) \geq \tau$ を満たす S．

4 局所探索法に基づくアルゴリズム

筆者らは，過去に配列設計問題に対するアルゴリズムを提案し，既存の研究より良い結果を得た [10, 11]．このアルゴリズムは局所探索法を基とし，組合せ的制約及び熱力学的制約を利用した場合の配列設計を行った．

提案手法では，局所探索法により制約を満たすよう配列集合の改善を行っている．単純な局所探索法では無く，排除連鎖法 [7, 19] と呼ばれる局所探索手法を適用した．また，移動戦略としては即時戦略を採用した．

4.1 近傍と評価関数

局所探索法を用いるには，近傍定義ならびに評価関数定義が必要となる．本節ではこれらを定義する，

近傍は以下のように定義する:

$$N(S) \stackrel{\text{def}}{=} \{S' \mid \text{sequence sets obtained by flipping1 base of a sequence belonging to } S\}. \quad (1)$$

3.1 節での制約を考える場合，評価関数は以下のように定義する:

$$\Delta G_{min}(S) \stackrel{\text{def}}{=} \min\{\Delta G_{ww}(S), \Delta G_{wc}(S), \Delta G_{cc}(S)\} \quad (2)$$

$\Delta G_{min}(S) \geq \tau$ の時，制約を満たしていることとなる．提案手法では，$\Delta G_{min}(S)$ の値を大きくしていく．

4.2 問題点

局所探索法を用いる際，改善解探索には近傍探索が必要となる．また，近傍解が改善解であるか否かの判定を行うためには，評価計算が必要となる．近傍解が改善解であるか否かの判定は繰り返し行う必要があるので，多くの評価計算が必要となる．しかしながら，$\Delta G(s, s')$ 計算には $O(n^3)$ 時間必要で，一度の評価計算に $O(m^2n^3)$ 必要となる．

近傍定義より，近傍操作前後の集合を比較するとき，変化のある配列は一本のみであるので，差分のみの計算を行えば，一度の評価計算に $O(mn^3)$ となる．

しかしながら，それでもこの計算時間は十分に大きいものであり，step 2 において，評価計算に必要な時間がネックとなってしまう．

5 提案手法

先述したように，評価計算に必要な時間が，アルゴリズム中においてネックとなっている．

そこで，評価計算を効率的に行うことで，高速化を目指す．以下に，二つのアイデアを述べていく．

5.1 MFE 計算における表の再利用

近傍定義より，近傍操作前後において，変化のある文字は一文字のみである．

MFE は DP を用いて計算される．この時，近傍操作前後で変化のある文字は一文字のみである為，DP で用いた表において，近傍操作前後で変化しない部分が存在する．なぜならば，図 6 で記したように，表のある値を決める際に必要な参照領域は，その値の右上に存在する部分のみ参照すれば良いためである．

よって，DP で用いる表を保持しておくことで，変化の無い部分の再計算は省略し，MFE 計算に必要な時間を軽減することができる．

5.2 近似計算の利用

局所探索法の特徴として，近傍中の多くの解は改善解でなく，改善解は少ししか存在しないことが一般的である．つまり，評価計算をするほとんどの場合は，改善解ではない近傍解に対する評価計算ということができる．

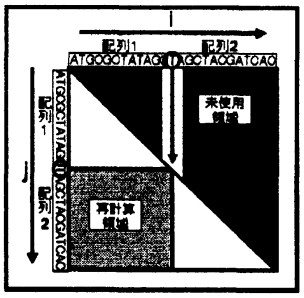

図 7: DP 表における近傍操作前後の差分

丸で囲んだ文字 (T) が近傍操作によって変化した文字とするとき，近傍操作前後で，表に変化の無い部分が存在する．

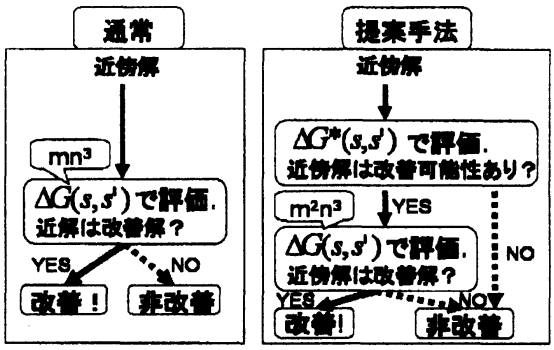

図 8: 近似 MFE の利用

もし，改善解でないものを $O(mn^3)$ 時間より小さな時間で判断できれば，評価計算にかかる時間全体の削減が図れると考えられる．そこで，明らかに改善解でないものを高速に判断するため手法について考えていく．

上記の事柄を実現するために，MFE の近似値を用いることを考える．ここで，近似 MFE を $\Delta G^*(s,s')$ と表記していくこととする．すると，MFE は「最小」自由エネルギーであるので，$\Delta G^*(s,s') \geq \Delta G(s,s')$ が明らかに成り立つ．この性質を利用して，明らかに改善解でないものを高速に判断する．

そのアイデアを記したのが，図 8 である．近傍探索中で解が改善解かそうでないかの判断を行う前に，$\Delta G^*(s,s')$ を用いて評価計算を行う．これにより，近似 MFE のより改善解でないと判断されたものは $\Delta G^*(s,s') \geq \Delta G(s,s')$ より，改善解の可能性がないと判断できる．よって，厳密な MFE 計算を行うことなく改善しない解を高速にはじくことができる．

この時，改善しない解を高速にはじく性能は $\Delta G^*(s,s')$ の性能に依存する．また，$\Delta G^*(s,s')$ の計算時間が遅いならば，このアイデアはうまく機能しない．そこで，近似 MFE 計算方法を提案する．

近傍定義より，局所探索法の近傍操作では塩基をひとつ置き換えることとしている．従って，近傍操作前後では，二本鎖での MFE となる構造も変化するが，塩基をひとつ置き換えただけなので，MFE となる構造の変化は小さいと仮定する．

そこで，近傍操作後の MFE 構造は，近傍操作前の MFE 構造と比較して，

1. 変化した塩基に関与していた塩基対が乖離
2. 変化した塩基が新たに塩基対を生成

したもののいずれかの構造に近いと判断する．そこで，上記のような構造のうち，最も小さなエネルギー値を $\Delta G^*_{neighbor}(s,s')$ とし，MFE の近似値とする．

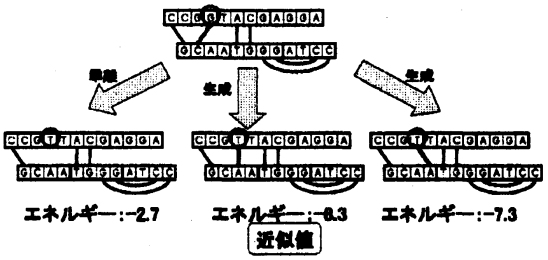

図 9: $\Delta G^*_{neighbor}$

5.3 予備実験

MFE 計算における表の再利用及び，$\Delta G^*_{neighbor}$ の性能を確認するために計算機

実験を行った.

100本の配列ペアをランダム生成し，一文字フリップで得られるもの全てに対してMFE及び近似MFEの計算を行った．MFE計算には`PairFold`パッケージ[2]を使用した.

結果は以下の通りである．ここで，通常のDPを用いてエネルギー計算を行ったものを「厳密DP」，5.1節のDP表の再利用適用は「省略DP」，5.2節の近似MFE$\Delta G^*_{neighbor}$適用は「neighbor」と記述している．また，各方法で計算したMFEもしくは近似MFEの値の平均を「計算値平均」(kcal/mol)，近似MFEの値と厳密DPの値の差の平均値を「差分平均値」(kcal/mol)，近似MFEの値と厳密DPの値の差の最大値を「差分最大値」(kcal/mol)，近似MFEの値と厳密DPの値が一致した回数を「一致数」(回)，各計算に要した時間を「計算時間」(sec)としている.

	厳密DP	省略DP	neighbor
計算値平均	-3.550	-3.550	-3.033
差分平均値			0.517
差分最大値			6.000
一致数			5504
計算時間	9.889	7.296	0.852

表 1: n=15, 全試行数:9000

	厳密DP	省略DP	neighbor
計算値平均	-4.866	-4.866	-4.237
差分平均値			0.629
差分最大値			6.400
一致数			6532
計算時間	38.230	26.778	1.332

表 2: n=20, 全試行数:12000

	厳密DP	省略DP	neighbor
計算値平均	-6.559	-6.559	-5.703
差分平均値			0.855
差分最大値			7.800
一致数			6164
計算時間	109.395	75.073	1.968

表 3: n=25, 全試行数:15000

省略DPでは，計算時間が厳密DPの75%程度に抑えられていることがわかる.

$\Delta G^*_{neighbor}$の性能を見てみると，計算時間は$n=15$の時で厳密DPの10分の1以下，$n=25$の時で厳密DPの50分の1以下と高速に近似可能なことがわかる.

6 計算機実験

以上の提案手法を実装し，計算機実験を行った．制約定数τを満たす，配列長n，集合サイズmのSを設計するのに必要な時間(sec)の測定を行った．この時，mとτの値が大きいほど，条件が厳しいと考える.

結果は以下の表である．ここで，「厳密DP」はDP表の再利用や近似MFEを使わずに，左のセッティングパラメータを満たす集合Sを設計するのにかかった時間(sec)を表している.「省略DP」はDP表の再利用を用いた場合，「省略DP+neighbor」はDP表の再利用と$\Delta G^*_{neighbor}$を用いた場合の時間(sec)を表している.

DP表の再利用した場合，使用しない場合と比較して，常に高速化できていることは明らかである.

$\Delta G^*_{neighbor}$を用いた場合を考える.

使用しない場合と比較して計算時間が大きくなっていることは無いが，セッティングパラメータによる条件が易しいとき(設計に要した時間が小さい時)は，あまり効果が出ていない．特に，$n=10, m=50, \tau=-4.0$の時が顕著である.

一方，条件が厳しいとき(設計に要した時間が大きい時)は，計算時間短縮に効果があることがわかる.

よって，$\Delta G^*_{neighbor}$は，条件が厳しい場合に有効であると判断できる．通常，配列設計問題では，条件が厳しい集合設計が求められる為，$\Delta G^*_{neighbor}$が条件が厳しい場合に有効であることは望ましい結果といえる.

7 まとめ

本予稿では，局所探索法による熱力学的制約を用いた塩基配列集合設計の改良法を提案し，計算機実験によりその有効性を示した.

n	m	τ	厳密 DP	省略 DP	省略 DP+ncighbor
10	50	-4.0	24.98	**22.88**	**22.88**
10	50	-3.5	84.16	74.72	**71.11**
10	100	-5.0	78.11	73.64	**73.43**
10	100	-4.0	738.47	662.26	**620.59**
15	50	-6.0	116.46	100.11	**86.52**
15	50	-5.0	576.98	486.70	**382.14**
15	100	-7.0	403.27	358.07	**315.12**
15	100	-6.0	1427.77	1209.02	**877.59**
20	50	-8.0	86.19	77.24	**76.97**
20	50	-7.0	163.67	142.44	**137.16**

表 4: 省略 DP・近似計算を利用した提案手法

参考文献

[1] L. Adleman: "Molecular Computation of Solutions to Combinatorial Problems", *Science*, Vol. 266(5187), pp. 1021–1024, 1994.

[2] M. Andronescu, Z. Zhang and A. Condon: "Secondary Structure Prediction of Interacting RNA Molecules", *J. of Molecular Biology*, Vol. 345(5), pp. 987–1001, 2005, Web page: `www.rnasoft.ca/download.html`.

[3] M. Arita, A. Nishikawa, M. Hagiya, K. Komiya, H. Gouzu and K. Sakamoto: "Improving Sequence Design for DNA Computing", *Proc. of 5th Genetic and Evolutionary Computation Conference*, pp. 875–882, 2000.

[4] M. Arita and S. Kobayashi: "DNA Sequence Design Using Templates", *New Generation Computing*, Vol. 20(3), pp. 263–273, 2002.

[5] Y. Asahiro: "Simple Greedy Methods for DNA Word Design", *Proc. of 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, Vol. 3, pp. 186–191, 2005.

[6] M. Garzon, V. Phan, S. Roy and A. Neel: "In Search of Optimal Codes for DNA Computing", *Proc. 12th DNA Computing*, LNCS(4287), pp. 143–156, 2006.

[7] F. Glover: "Ejection Chains, Reference Structures and Alternating Path Methods for Traveling Salesman Problems", *Discrete Applied Mathematics*, vol. 65(1–3), pp. 223–253, 1996.

[8] S. Kashiwamura, A. Kameda, M, Yamamoto and A. Ouchi: "Two-Step Search for DNA Sequence Design", *Proc. of the 2003 International Technical Conference on Circuits/Systems, Computers and Communications*, pp. 1889–1892, 2003.

[9] S. Kashiwamura, M. Yamamoto, A. Kameda, T. Shiba and A.Ouchi: "Potential for Enlarging DNA Memory: The Validity of Experimental Operations of Scaled-up Nested Primer Molecular Memory", *BioSystems*, Vol. 80(1), pp. 99–112, 2005.

[10] S. Kawashimo, H. Ono, K. Sadakane and M. Yamashita: "DNA Sequence Design by Dynamic Neighborhood Searches", *Proc. of 12th DNA Computing*, LNCS(4287), pp. 157–171, 2006.

[11] S. Kawashimo, H. Ono, K. Sadakane and M. Yamashita: "Dynamic Neighborhood Searches for Thermodynamically Designing DNA Sequence", *Proc. of 13th DNA Computing*, LNCS(4848), pp. 130–139, 2008.

[12] S .Kobayashi, T. Kondo and M. Arita: "On Template Method for DNA Sequence Design", *Proc. of 8th DNA Computing*, LNCS(2568), pp. 205–214, 2002.

[13] R. Lyngsø, M. Zuker and C. Pedersen: "Fast evaluation of internal loops in RNA secondary structure prediction", *Bioinfomatics*, Vol. 15, pp. 440–445, 1999.

[14] M. Shorteed, S. Chang, D. Hong, M. Phillips, B. Campion, D. Tulpan, M. Andronescu, A. Condon, H. Hoos and L. Smith: "A thermodynamic approach to designing struct-free combinatorial DNA word set", *Nucleic Acids Research*, Vol. 33(15), pp. 4965–4977, 2005.

[15] F. Tanaka, A. Kameda, M. Yamamoto and A. Ohuchi: "Design of nucleic acid sequences for DNA computing based on a thermodynamic approach", *Nucleic Acids Research*, Vol. 33(3), pp. 903–911, 2005.

[16] D. Tulpan, H. Hoos and A. Condon: "Stochastic Local Search Algorithms for DNA Word Design", *Proc. of 8th DNA Computing*, LNCS(2568), pp. 229-241, 2003.

[17] D. Tulpan and H. Hoos: "Hybrid Randomized Neighborhoods Improve Stochastic Local Search for DNA Code Design", *Proc. Advances in Artificial Intelligence, 16th Conference of the Canadian Society for Computational Studies of Intelligence*, LNCS(2671), pp. 418–433, 2003.

[18] D. Tulpan, M. Andronescu, S. Changf, M. Shortreed, A. Condon, H. Hoos and L. Smith: "Thermodynamically based DNA strand design", *Nucleic Acids Research*, Vol. 33(15), pp. 4951–4964, 2005.

[19] M. Yagiura, T. Ibaraki and F. Glover: "An Ejection Chains Approach for the Generalized Assignment Problem", *INFORMS Journal on Computing*, Vol. 16(2), pp. 133–151, 2004.

[20] M. Zuker and P. Stiegler: "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information", *Nucleic Acids Research*, Vol. 9, pp. 133–148, 1981.